Publication of science data products through the ESO archive: lessons learned and future evolution

Jörg Retzlaff*^a, Magda Arnaboldi^a, Nausicaa A. R. Delmotte^a, Laura Mascetti^b, Alberto Micol^a ^aEuropean Southern Observatory, Karl-Schwarzschild-Str. 2, D-85748 Garching, Germany; ^bTerma GmbH, Darmstadt, Germany

ABSTRACT

Phase 3 denotes the process of preparation, submission, validation and ingestion of science data products for storage in the ESO Science Archive Facility and subsequent publication to the scientific community. In this paper we will review more than four years of Phase 3 operations at ESO and we will discuss the future evolution of the Phase 3 system.

Keywords: Data products, astronomical archives, surveys, data access, observatory operations

1. INTRODUCTION

The European Southern Observatory Science Archive Facility (SAF) evolved from an archive containing predominantly raw data into a resource also offering science-grade data products ready for immediate scientific interpretation or further analysis steps that may be required to address a specific scientific question¹.

New products originate from two different sources. On the one hand Principal Investigators of Public Surveys and other programmes reduce the raw observational data and return their products using the so-called Phase 3 - a process that extends the Data Flow System after proposal submission (Phase 1) and detailed specification of the observations (Phase 2). On the other hand raw data of selected instruments and modes are uniformly processed in-house, independently of the science goal for which the data had been originally acquired. Making ESO data products in a timely and operationally sustainable fashion available to the community represents the primary objective of Phase 3.

ESO Public Surveys are large observing programmes with a substantial share in telescope time targeting fundamental questions in astrophysics today. To ensure their legacy value, the survey teams are committed to return the reduced data for publication to the community².

Currently, data products assets in the ESO science archive facility consist of calibrated images and spectra, as well as catalogues, mostly sampling the optical-NIR bands of the electromagnetic spectrum beside including some sub-millimetre data.

Data validation prior to archive ingestion represents an essential part of the standard Phase 3 procedures. Verifying the compliance to the standard data format, the existence of standard keywords for data description, and the consistency with the associated documentation ensures science-driven data discoverability, uniform description of and access to archive assets independent of the data reduction per se.

Both the data products contributed by the community and those generated at ESO are in great demand by science archive users. From the first publication of data products in 2011 to January 2016, in excess of 1300 unique users have accessed products of either origin³.

^{*}joerg.retzlaff@eso.org phone +49 89 3200 6729; fax +49 89 3200 6358; www.eso.org

2. OVERVIEW OF THE PHASE 3 PROCESS

In ESO jargon "Phase 3" denotes the process of preparation, validation and ingestion of science data products for storage in the SAF, and subsequent data publication to the scientific community.

The main point of access for Principal Investigators and their collaborators who return reduced data products resulting from ESO observations is the so-called Phase 3 Release Manager – a dedicated web application for controlling the entire Phase 3 data submission process (Fig. 1). Initially, the system prompts the user to log in using the personal ESO User Portal credentials.

1. Registration of Phase 3 submission

Phase 3 submissions are associated with ESO observing programmes. If the respective programme is not listed under *Data Collections* or if the display is empty, the Phase 3 submission need to be registered by sending an e-mail message to <u>USD-HELP@ESO.ORG</u> quoting the subject "REQUEST FOR PHASE 3 PROGRAMME <PPP.C-NNNN>", where <PPP.C-NNNN> denotes the ESO programme identifier.

Thereafter, usually within one working day, the confirmation that the programme has been registered will be sent by email and the entry will show up in the Phase 3 Release Manager as a new *Data Collection*.

2. Data Preparation

Data should be prepared according to the data formats specified in the ESO Science Data Products Standard: ESO/SDP Standard (v5), where applicable taking into account the specifications for one-dimensional spectra⁴, APEX Sub-mm Sky Maps, Integral Field Spectroscopy (3D Data Cubes), ESO/GTC data products (spectra and images) and the amendments published in the FAQ section of the Phase 3 web pages.

3. Uploading data to ESO

Any FTP client may be used to transfer the data to the dedicated Phase 3 FTP server

ftp://phase3ftp.eso.org/<name>/batch ###

using the directory path displayed in the Phase 3 Release Manager. The path includes the <name> of the Phase 3 Data Collection and the Batch ID.

The directory and the data deposited therein are visible and accessible only to the P.I. who owns the Phase 3 collection, the survey manager and appointed delegates. In the course of preparing the data release you can modify files on the FTP server as needed, including file replacements and removal.

Possible problems concerning the file format need to be resolved first and the corrected files must be uploaded before proceeding.

4. Verifying compliance with data format requirements

Once having completed the data transfer, the submission need to be 'closed' in order to proceed (by pressing the CLOSE button in the Phase 3 Release Manager).

This signal triggers the *Phase 3 format and provenance verification* process on the ESO server, which requires time between a couple of minutes and several hours to complete depending on the amount of data⁵.

The user will receive an e-mail notification upon completion of the process. If problems are reported at this stage, these need to be fixed before proceeding. In this case the directory in the Phase3ftp server is still write-mode enabled.

Note: "closing" the Phase 3 batch turns the respective FTP directory into read-only mode and one cannot upload further files, nor replace or delete previously uploaded files.

5. Uploading the data release description

In this step the data release description document associated with the uploaded Phase 3 data must be uploaded (PDF format) to provide an account of the release content, the originating observations, the calibration and data reduction procedures, the data quality, the data format, and, possibly, the scientific context of the programme.

6. Finalising the data submission

Before the data are made publicly available to the community at large, the "SUBMIT" button need to be pressed to confirm the completeness and consistency of the actually uploaded data with respect to the data release description.

To this end the Phase 3 Release Manager application provides a detailed summary of the submitted data in terms of content, i.e. total number of files, data types, date of observation, sky coverage etc. ("Content Summary").

Note: The final data submission must be done by the P.I., this action cannot be delegated.

Thereafter, the submitted data and documentation will undergo a final check by ESO, dubbed *Phase 3 content validation*, before being published through the ESO Science Archive Facility.

In case any problems are spotted at this stage, the user will receive a detailed report including instructions of how to proceed.

Finally, as result of the successful Phase 3 submission, the new data release will be added to the overview table of available releases (Fig. 2) and the data will appear in the ESO/SAF archive query interfaces for searching and download.

3. EVOLUTION OF PHASE 3 OPERATIONS

A heterogeneous suite of software components running at ESO headquarters in Garching supports the Phase 3 workflow. The original software, which was deployed in March 2011, had been designed before the first ESO public surveys officially started operations on the ESO/VISTA survey telescope in April 2010, that is more than 6 years ago. In the meantime the publication of science data products via Phase 3 has been ramping up continuously, exhibiting a rate of increase that pushed the system capabilities to its operational limits.

List of selected milestones marking the steady development of Phase 3 operations (see http://www.eso.org/sci/observing/phase3/news.html for a complete list)

- March 2011 Start of Phase 3 operations to support the first data release of ESO/VISTA public surveys.
- October 2012 Added support for the data submission, validation, querying, and access of very large photometric catalogues being produced by public surveys.
- October 2012 To support ESO spectroscopic surveys in their Phase 3 preparations the format standard for data products has been expanded and now includes spectroscopic data, i.e. extracted, calibrated, and fully characterized one-dimensional spectra.
- October 2013 Publication of one-dimensional calibrated spectra of point-like sources obtained by the <u>UVES</u> instrument the cross-dispersed echelle spectrograph mounted at UT2 of the Very Large Telescope (VLT), based on the homogeneous pipeline processing of archival data and the stream of newly acquired data.
 - May 2014 Access opened to 33000 reduced science-ready spectra obtained with the XSHOOTER spectrograph mounted on the VLT.
- October 2014 HARPS (High Accuracy Radial Velocity Planet Searcher mounted on the 3.6-m telescope at the ESO La Silla) pipeline processed data available through Phase 3
 - March 2015 Phase 3 FTP storage extension deployed including an alignment of the operational architecture.
 - May 2015 Release of pipeline-processed Flames/Giraffe one-dimensional spectra through the ESO Science Archive Facility.
- December 2015 Added support for 3-dimensional data cubes (integral field spectroscopy).
 - January 2016 870-µm emission maps and source catalogues produced by the ATLASGAL APEX survey published via Phase 3.

Combining the feedback received from Phase 3 users with our own experience of operating the Phase 3 system accumulated during the past years led to a prioritized list of improvements and features to be implemented.

In view to the increasing operational load due to upcoming Phase 3 data releases for thirteen different public survey projects, data submissions by other programmes and in-house pipeline data processing, and taking into account the size and complexity of the currently available data releases (Fig. 2), a major upgrade of the Phase 3 software took place in 2016 Q2.

4. PHASE 3 UPGRADE

Further streamlining of Phase 3 operations, both from the user's and the operator's perspective, is the primary goal of the upgrade, which will be achieved by adding a number of essential features that were missing thus far. For instance, to reduce the overall time to publication, shortly after having uploaded the data, before in-depth content validation takes place, the user gets notified in a timely fashion if mandatory FITS header keywords are missing.

Native support for very large releases through multi-batch submission scheme

The *Phase 3 Batch* concept has been introduced in addition to the concept of *Phase 3 Data Release* to allow the submission and publication of very large releases in terms of multiple chunks of data being submitted sequentially or concurrently depending on the needs of the data provider.

In this way very large data releases can be split into a sequence of smaller batches, which can be managed according to the data provider's preferences independently from the data release at large.

The new flexibility also allows survey teams to work on the submission of different data product releases, e.g. single-band products on the one hand side and band-merged catalogs on the other hand in a parallel fashion.

This implementation ensures that the data validation can be successfully carried out independently of scalability issues. Previous experience showed that there is a hard limit of ca. 2 TB in the total volume. This conservative approach is particularly useful to manage the next big VISTA submissions such as the complete VVV survey data release covering about six years of observations.

Test Mode

It is possible to verify the compliance of a *test data set* with data format requirements without actually submitting those data for archiving and publication, i.e. without going through the entire Phase 3 data submission work flow (choosing "TEST YOUR DATA" from the action menu within the Phase 3 RM). In this way the *Phase 3 format and provenance verification* process will be executed, after which the batch directory will be reset to read/write mode, i.e. to the "OPEN" state.

The user will receive a notification via e-mail upon completion of the validation process and the resulting test report will be accessible from the RM.

Phase 3 survey manager role

With the goal to improve the interface between public survey teams and ESO during Phase 3 activities, the new role of "Phase 3 survey manager" has been introduced. Before starting a Phase 3 submission, the P.I. of the survey programme must assign the survey manager role to her/himself or another team member.

The survey manager has the overall responsibly for delivering Phase 3 data products according to the originally stipulated survey management plan. The survey manager has the same permissions as the P.I. including Phase 3 delegation and release submission, except for re-assigning the survey manager role.

Updating previously released Phase 3 data

Previously released Phase 3 data can be updated by submitting a new Phase 3 batch. The following use cases are supported:

- consecutive updating data releases with increasing signal-to-noise ratio in the final combined data product may occur according to the progress of acquiring more and more observational data in the course of a programme aiming at very deep integrations on the same target. This happens typically when observations are scheduled over more than one period;
- reprocessing of previously released data motivated by improved data calibration procedures, improved quality control checks or enhanced data reduction algorithms, leading to a higher data quality than before.

To identifying old with new files, depending on the adopted naming convention for the new files, there are two different options to provide the required link between new (updating) and old (to be updated) files.

a) CHANGES.USER. Each line in the CHANGES.USER special file must follow the pattern

<new_file> UPDATES <old_file>

with <new_file> referring to a file in this batch submission, and <old_file> referring to a previously submitted file within the same data collection/survey (both in terms of the originally submitted filename, also known as ORIGFILE).

b) Automatic tracking of data versions. If a new file is submitted with the same filename (ORIGFILE) as the earlier version, then these two files are automatically recognized by the system as different data versions without a line in CHANGES.USER being needed for this file.

Within one and the same Phase 3 batch CHANGES.USER and automatic tracking cannot be used simultaneously.

Updating procedure: Though the submission for updating Phase 3 batches basically follows the same procedure as for batches containing new data, the following points deserve particular attention.

- Improved versions of previously submitted files must be uploaded to a new Phase 3 batch within an existing data collection.
- Unless automatic version tracking is used (see above) the CHANGES.USER file need to be uploaded. Note: one can upload CHANGES.USER via FTP or via the Release Manager.
- Data release description: for updating releases a dedicated section documenting changes and improvements of the new data with respect to the old data is required.
- Important check for updates: when reviewing the *completeness* and *consistency* of the Phase 3 batch prior to submission, users should check that the *update status* is correct ("UPDATING") and the respective file counts are correct.

Result: Once the updating batch has been published the new version of the data becomes available, replacing the old obsolete version in any default archive query. Obsolete data are only accessible on demand.

NEW+UPDATING: updates can be combined with the submission of new data, i.e. data products which have not been submitted via Phase 3 yet, within the same Phase 3 batch. The automatic tracking of data versions implies that previous data cannot be referenced in terms of ORIGFILE.

New header update mechanism for science files

To update the astrometric or photometric calibration of a previously released FITS image it is now possible to submit just the modified portion of the FITS header information instead of the entire FITS file, thereby drastically reducing the total amount of data to be transferred in such cases.

Reporting a Data Problem

Users of the ESO Science Archive Facility are generally encouraged to provide feedback regarding the data quality of the downloaded data products by contacting the ESO Archive Science Group via usd-help@eso.org, subject: Phase 3.

As *Phase 3 data provider*, despite exercising due diligence during data preparation and submission, one may notice only after publication through the ESO science archive, that some previously provided Phase 3 data actually do not fulfill the criteria for being released as science data products and that these data should be flagged accordingly in the user interface to notify archive users when selecting data.

In such case the user need to go to the RM and click "REPORT DATA PROBLEM" to specify the collection containing the data in question, the kind of problem, and the list of files being affected.

The best matching description for the kind of problem need to be selected from the available list, or, if the problem type is not listed, the user should enter her/his own description.

Default list of data problems:

Data quality: data corrupted

Data quality: high background

Data quality: highly variable background

Data quality: telescope tracking error

Data quality: below quality cuts

Data quality: astrom. calibration failure

· Data quality: flux calibration failure

Data quality: wavelength calibration failure

Data quality: other calibration failure

Data content: technical/calibration data, not science

Data content: does not belong to program

· Data content: wrong pointing

Data content: spurious duplication

· Header keywords: RA/DEC incorrect

Header keywords: non-compliant

Data problems are reported per "*Phase 3 data set*", i.e. the filename of the science file (having PRODCATG=SCIENCE.*) must be listed and the ancillary ('associated') files, if any, will be automatically included.

The ESO Archive Science Group is responsible for handling the report. Having validated the input, the respective data sets will be marked as "deprecated" in the ESO archive, which entails that these data are not available by default but only on specific request.

The P.I., equivalently, the owner of the Phase 3 collection, or the Phase 3 survey manager, are entitled to report data problems.

5. CONCLUSIONS

Successfully ingesting and publishing reduced data during the past five years via Phase 3 relies on the following proven and tested pillar concepts.

- 1. the ESO/SDP data format standard defining a unique format for each data type including a minimal common set of keywords for data description;
- 2. the data release description forms an integral part of each Phase 3 release, providing a human readable description of the data content in terms of a user manual;
- 3. Phase 3 data validation implementing a standardized procedure within the Phase 3 workflow to assure the overall quality of each release by verifying completeness and consistency of the data content, associated metadata information and documentation [6].

The Phase 3 software upgrade, which was deployed in 2016 Q2, implements a series of improvements in order to reduce the overall time for data publication. Now, the system provides swift feedback regarding the compliance of uploaded data with respect to the required data format specification. An initial check takes place immediately after having uploaded a new file andthe referential integrity of the submitted provenance pointers is checked after 'closing' the batch.

The new "Phase 3 Content Summary" provides additional feedback about the uploaded data content in terms of aggregate information (e.g. number of files, sky coverage in square degrees) broken up according to various key parameters like time of observation, filter band, image quality etc.

Further improvements consist of implementing standard tasks in software instead of requiring manual interventions, which involve different operational units for resolution (e.g. management of batches and releases).

Currently, the 4th data submission for the ESO/VISTA public surveys in ongoing using the new Phase 3 software with almost 17 TB of new imaging data being uploaded thus far. Apart from some reliability issues, which are currently under investigation, the initial feedback from the Phase 3 users about new capabilities of the upgraded system are very positive.

ACKNOWLEDGEMENTS

This work is largely based on the contributions of reduced data by ESO Phase 3 users. We are particular grateful for the constructive feedback and suggestions for improvements of the Phase 3 system received. Software maintenance/development and IT support for the Phase 3 project as provided by our ESO colleagues in the Directorate for Engineering are warmly acknowledged. Stefano Zampieri managed the development activities, Vincenzo Forchi, Malgorzata Stellert, Fabio Sogni, Ignacio Vera Sequeiros implemented the software, Maurice Klein Gebbinck and Moreno Pasquato were responsible for testing, Livio Condorelli and Zsolt Zsiros helped setting up the IT infrastructure.

REFERENCES

- [1] Retzlaff J. et al., Proceedings of the SPIE, Volume 9149, id. 914903 8 pp. (2014)
- [2] Arnaboldi M. et al., Paper 9910-11, this volume.
- [3] Romaniello M. et al., The Messenger 163, March 2016
- [4] Micol A. et al., Paper 9910-121, this volume.
- [5] Mascetti L. et al., Paper 9910-87, this volume.
- [6] Delmotte N. et al., Paper 9910-7, this volume.

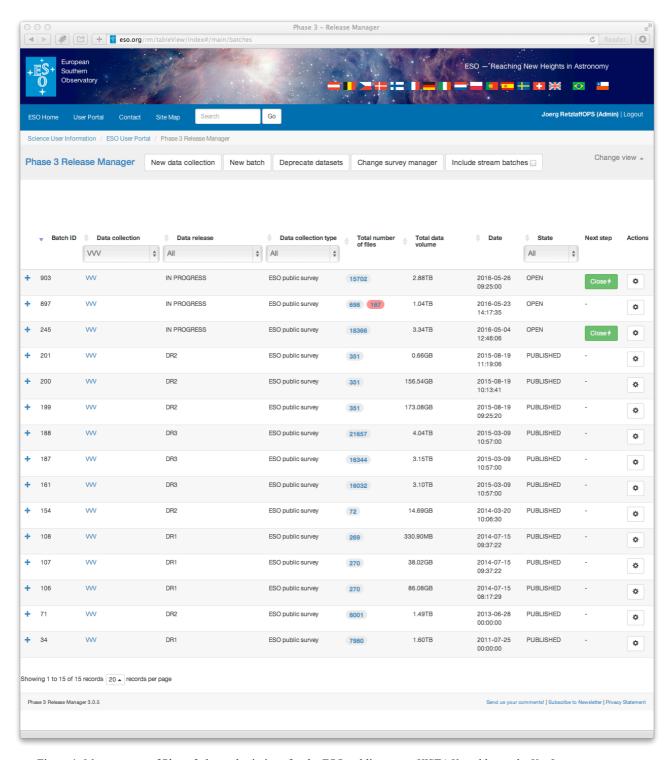


Figure 1: Management of Phase 3 data submissions for the ESO public survey VISTA Variables in the Via Lactea.

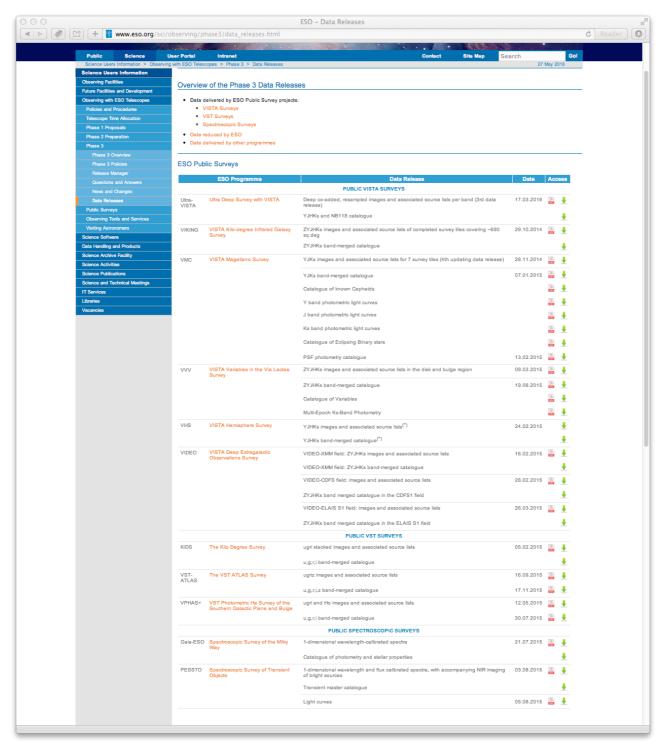


Figure 2: Overview of ESO public survey data releases.